LGEM$^+$: Automated Improvement of Metabolic Network Models and Model-Driven Experimental Design through Abduction

Alexander H. Gower[1*], Konstantin Korovin[2], Daniel Brunnsåker[1], Erik Y. Bjurström[1], Praphapan Lasin[1], Ievgeniia A. Tiukova[1,3], Ross D. King[1,4,5]

[1*]Chalmers University of Technology, Gothenburg, Sweden.
[2]The University of Manchester, Manchester, United Kingdom.
[3]KTH Royal Institute of Technology, Stockholm, Sweden.
[4]Cambridge University, Cambridge, United Kingdom.
[5]Alan Turing Institute, London, United Kingdom.

*Corresponding author(s). E-mail(s): gower@chalmers.se;
Contributing authors: Konstantin.Korovin@manchester.ac.uk; danbru@chalmers.se; erikbju@chalmers.se; lasin@chalmers.se; tiukova@chalmers.se; rossk@chalmers.se;

Abstract

Scientific discovery in biology is difficult due to the complexity of the systems involved and the expense of obtaining high quality experimental data. Automated techniques are a promising way to make scientific discoveries at the scale and pace required to model large biological systems. A key problem for 21st century biology is to build a computational model of the eukaryotic cell. The yeast *Saccharomyces cerevisiae* is the best understood eukaryote, and genome-scale metabolic models (GEMs) are rich sources of background knowledge that we can use as a basis for automated inference and investigation.

We present LGEM$^+$, a system for automated abductive improvement of GEMs, and for experimental design consisting of: a compartmentalised first-order logic framework for describing biochemical pathways (using curated GEMs as the expert knowledge source); a two-stage hypothesis abduction procedure; integration with flux-balance analysis (FBA); and metabolic pathway extraction and analysis algorithms.

We demonstrate that deductive inference on logical theories created using LGEM$^+$, using the automated theorem prover iProver, can predict growth/no-growth of *S. cerevisiae* strains in minimal media. LGEM$^+$ proposed 2094 unique candidate hypotheses for model improvement. We assess the value of the generated hypotheses using two criteria: (a) genome-wide single-gene essentiality prediction, and (b) constraint of flux-balance analysis (FBA) simulations. We present a model-driven experimental design strategy, and demonstrate this with a differential expression study, and using the **Δ***pfk2* mutant strain as a case study.

Keywords: Scientific discovery, artificial intelligence, systems biology, metabolic modelling, first-order logic, automated theorem proving.

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Alice Wallenberg Foundation. Funding was also provided by the Chalmers AI Research Centre and the UK Engineering and Physical Sciences Research Council (EPSRC) grant nos: EP/R022925/2 and EP/W004801/1, as well as the Swedish Research Council Formas (2020-01690).

1 Introduction

An important aspect of systems biology is improving our understanding of cellular processes, and the complex interactions between genes, proteins and chemical species. Systems biology is the research discipline that tackles this complexity. *Saccharomyces cerevisiae*, commonly known as "baker's yeast", is an excellent model organism used for the study of eukaryote systems biology. This is due to the availability of tools for easy genetic manipulation, and low cultivation cost, enabling targeted experiments to characterise the system. *S. cerevisiae*'s was the first eukaryotic genome to be fully sequenced (Goffeau et al, 1996) and there is a wealth of knowledge about the gene functions, many of which are conserved or expected to have equivalents in other eukaryotes, including humans (Dujon, 2010). Metabolic network models (MNMs) represent the cellular biochemistry of an organism and the related action of genes encoding enzymes; such models which seek to integrate knowledge from the entire organism are known as genome-scale metabolic models (GEMs).

The purposes of LGEM$^+$ are:

P1 to formalise existing knowledge about *S. cerevisiae* in a logical framework;
P2 to use this knowledge to, for varying conditions and genotype, efficiently predict phenotype; and
P3 to pinpoint missing knowledge, and suggest ways to amend this knowledge via a suitable hypothesis (abduction) that can be experimentally validated. LGEM$^+$ is designed not only for prediction, but for scientific discovery: to improve upon draft models, specifically GEMs.

These three purposes follow closely the high-level framework for improvement of GEMs outlined by Thiele and Palsson (2010), consisting broadly of three stages: hypothesise refinements to the model; convert the refined model to a format suitable for simulation; and evaluate based on experimental evidence and internal consistency. Evaluation is dependent on executing simulations using a mathematical formalism. However optimising a model for a specific mathematical formalism is not the scientific goal—any improvements that are made to a GEM within a certain framework should translate to improvements in the underlying knowledge.

Model quality in GEMs is multi-faceted—examples of desirable properties of a model are: predictive power; metabolic network coverage; and parsimony. There are trade-offs between different desirable properties (Heavner and Price, 2015). Foremost, however, is the predictive power of the GEM. Ultimately the aim is to understand the entities, mechanisms and adaptations that govern yeast growth in different environments. Challenges for the future of genome-scale modelling of *S. cerevisiae* include: improving annotation; removing noise from low-confidence components; and adding reactions to eliminate so-called "dead-end" compounds (Chen et al, 2022). To multiply the efforts of human researchers, previous work has investigated automating parts of the scientific method. GrowMatch was a technique developed to resolve inconsistencies between predictions and experimental observations of single-gene mutant strains of *Escherichia coli* (Kumar and Maranas, 2009). Other approaches to metabolic network gap-filling have exploited answer-set programming, for example MENECO which is designed to efficiently identify candidate additions to draft network models (Prigent et al, 2017). Other methods for computational approaches to revising metabolic network models are reviewed in Tamura et al (2015).

Logical inference can be applied to generate and improve metabolic models: induction allows us to generalise models from data; given a theory we can draw conclusions using deduction; and abduction enables us to form hypotheses to improve consistency with empirical data. In this work we use first-order logic (FOL) to simulate the metabolic network, an approach first proposed in 2001 (Reiser et al, 2001). A FOL model was used to generate functional genomics hypotheses then tested by a robot scientist (King et al, 2004); logical induction and abduction was applied to identify inhibition in metabolic pathways after introduction of toxins (Tamaddoni-Nezhad et al, 2005); and an FOL model constructed in Prolog using the GEM iFF708 (Förster et al, 2003) as the background knowledge source was used to predict single-gene essentiality (Whelan and King, 2008). Huginn is a tool that uses abductive logic programming (ALP), and demonstrates the ability to improve metabolic models and suggest *in vivo* experiments (Rozanski et al, 2015).

A key distinction of our model compared to bespoke algorithmic methods such as MENECO, is our use of automated theorem provers for first-order logic (ATPs) to perform deductive and abductive inference. This approach shifts the burden of abductive algorithm design away from domain-specific implementation. Previous FOL approaches have used Prolog, but we decided to work with ATPs. For the reasoning tasks we use the ATP iProver (Korovin, 2008), which was chosen due to its performance and scalability as well as completeness for first-order theorem finding. We extended iProver to include abduction inference. iProver is a saturation-based

theorem prover that saturates via consequence finding algorithms which are well-suited to abduction (Simon and del Val, 2001). Using ATPs will also allow us to combine different deduction and abduction strategies.

Furthermore, our model is capable of deductive and abductive reasoning at scales far greater than previous FOL approaches. The ability to reason at scale is particularly important for the automation of scientific discovery in systems biology where the domain is complex, and empirical data coverage is sparse, and expensive to generate.

The FOL framework in LGEM$^+$ does not currently include information on reaction stoichiometry. Therefore to integrate quantitative modelling, we propose in this paper a method to combine flux balance analysis (FBA) and logical inference to validate metabolic pathway configurations found by LGEM$^+$.

This work extends our previous work (Gower et al, 2023) with several novelties incorporated into the modelling framework, LGEM$^+$ and by applying the system to the task of experimental design, demonstrated with a case study. In particular, the novel areas of the method are as follows.

- We developed a new pathway extraction algorithm to compare against both FBA simulations and expected pathways.
- We incorporated knowledge on compound relationships to abduce fixes to a *S. cerevisiae* GEM.
- We exploited our algorithms to design specific experiments.
- We predicted changes to the central carbon metabolism under single-gene deletions, and analyse the prediction for $\Delta pfk2$.
- We compared experimental data against predictions from the model.

The main contributions of LGEM$^+$ as presented in this paper are: (1) a compartmentalised FOL model of yeast metabolism; (2) a set of algorithms for the extraction and analysis of metabolic pathways from simulations; (3) a two-stage method for the abduction of novel hypotheses on improved models; (4) scalable methods for evaluating these models and hypotheses, with parallelised implementation; (5) an algorithm to integrate FBA with abductive reasoning.

2 Methods

2.1 The First-order Logic Framework

We chose FOL as the language to express the mechanism of the biochemical pathways. FOL allows for a rich expression of knowledge about biological processes, such as reactions and enzyme catalysis. We use FOL to express our knowledge about how entities are known to interact, for example that a reaction has substrates and products, and possibly some annotated enzyme. The method and model we design is independent of the specific network, meaning that although here we apply LGEM$^+$ to *S. cerevisiae*, this modelling framework could equally well be applied to other organisms.

We define five predicates in the first-order language: met/2, gn/1, pro/1, enz/1, and rxn/1. The semantic interpretation of these predicates is outlined in Table 1. Here a cellular "compartment" refers to a component of the cellular anatomy, e.g. mitochondrion, nucleus or cytoplasm. The specific compartments vary depending on the GEM, for example the compartments included in Yeast8 (Lu et al, 2019) are: cell

Table 1 Predicates used in the logical theory of yeast metabolism. Forward and reverse reactions are represented separately in the model, thus a "positive flux" through a reversed reaction indicates the reaction flux is negative.

Predicate	Arguments	Natural language interpretation
met/2	metabolite, compartment	"Metabolite X is present in cellular compartment Y."
gn/1	gene identifier	"Gene X is expressed."
pro/1	protein complex identifier	"Protein complex X is available (in every cellular compartment)."
enz/1	enzyme category identifier	"Enzyme category X is available."
rxn/1	reaction	"There is positive flux through reaction X."

envelope, cytoplasm, extracellular, mitochondrion, nucleus, peroxisome, endoplasmic reticulum, golgi, lipid particle, vacuole, endoplasmic reticulum membrane, vacuolar membrane, golgi membrane, and mitochondrial membrane.

Clauses in our model are one of seven types, each expressing relationships between entities in terms of the predicates given above. These components of LGEM$^+$ are listed below, and we provide a graphical overview and example statements in Fig. 1.

C1 **Reaction activation** clauses state that all substrate compounds for a specific reaction being present in the correct compartments, together with availability of a relevant enzyme, implies the reaction is active.
C2 **Reaction product** clauses state that a reaction being active implies the presence of a product compound in a given compartment.
C3 **Enzyme availability** clauses state that the availability of the constituent parts (proteins) of an enzyme imply the availability of the enzyme. Enzymes sometimes act in complexes made up of two or more proteins, and different enzymes that catalyse the same reaction are called isoenzymes.
C4 **Protein formation** clauses state that the presence in the genome of a gene that codes for a specific protein implies the availability of that protein.
C5 **Gene presence** clauses are statements expressing either the presence or absence of a particular gene in the genome.
C6 **Metabolite presence** clauses are statements expressing the presence of a particular compound in a specific compartment.
C7 **Goal** clauses represent a biological objective, usually the presence in the cytoplasm of a set of compounds deemed essential for growth, but could also be another pathway endpoint or intermediary compound.

LGEM$^+$ can be used for the following tasks: deductions to assess growth and production of compounds; predicting gene essentiality; abduction of hypotheses on the model; predicting possible pathways and assessing their feasibility with FBA; and selecting experiments value. We will now cover the methods for each of these tasks in more detail.

2.2 Assessing Growth and Production of Compounds

Yeast growth is dependent on the production of essential chemical products—intermediary points or endpoints of biochemical pathways within the organism. The core of these biochemical pathways is the enzymatic reactions, and they are facilitated

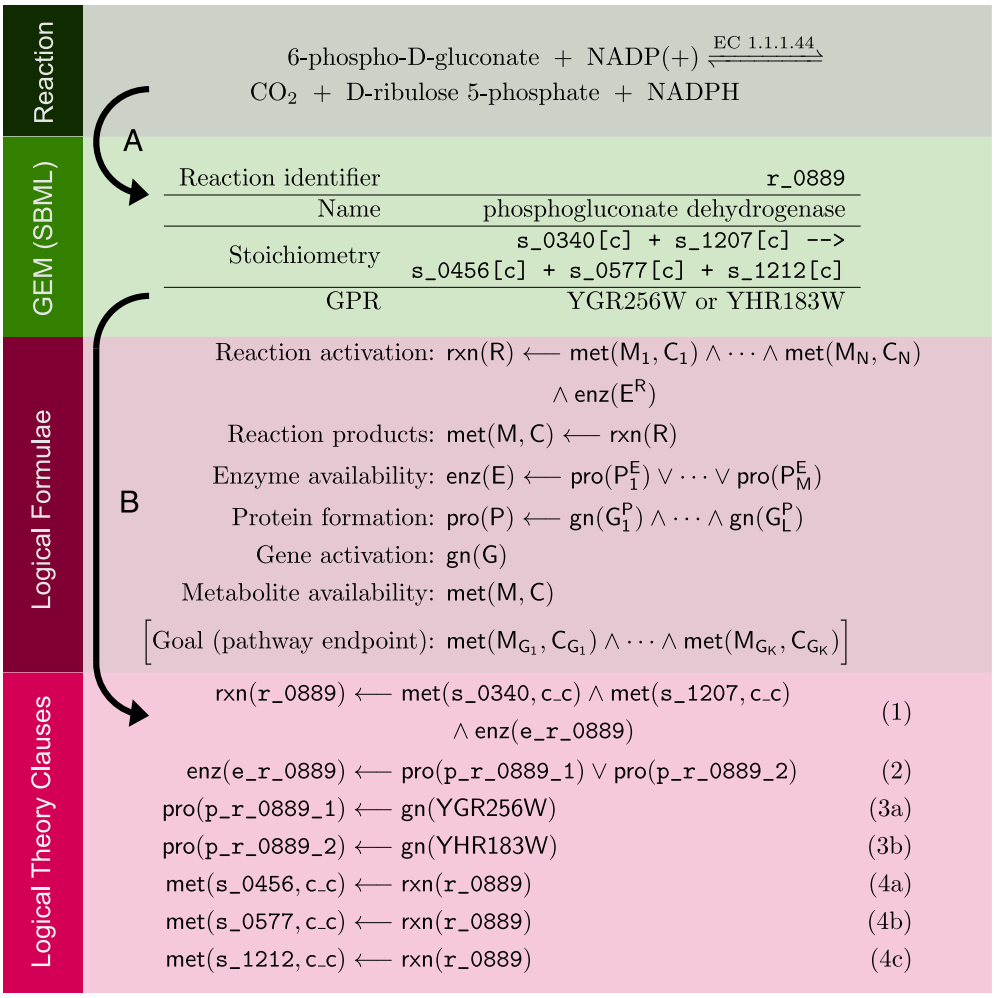

Fig. 1 Conversion of genome-scale metabolic model provided in Systems Biology Markup Language (SBML) to logical theory. **(A)** A reaction is encoded in SBML using identifiers to represent the substrates and products, and a logical rule for enzyme availability (GPR="gene-protein-reaction rule"). **(B)** The information contained on each reaction is encoded using logical formulae into a set of clauses; predicate definitions are provided in Table 1. Here equation (1) is the reaction activation clause. "∧" is a conjunction symbol ("AND"), meaning all of the literals in the expression must be true for the RHS of the clause to be true; "∨" is a disjunction symbol ("OR"). So we can read (1) as: "reaction `r_0889` is active if all of the metabolites in the set {`s_0340`, `s_1207`} are present in the cytoplasm and at least one of the isoenzymes is present". Similarly equation (2) describes the condition for a relevant enzyme to be present; equations (3a,b) describe the conditions for each of these isoenzymes to be formed; and equations (4a-c) are the reaction product clauses and state that "if reaction `r_0889` is active then each of its products is present".

by transport via transporters across compartment boundaries or the cell membrane, and diffusion of chemicals within cellular compartments, including the cytoplasm. Certain products are deemed essential for growth, so if production of these compounds is inhibited then the organism has no viability.

LGEM$^+$ can be used for growth simulations in the following way. The general formulation of the problem provided to the ATP is to identify whether a theory, T, "entails" a goal, G. In other words that the goal is a logical consequence of the theory ($T \models G$). Here T is a set of logical axioms that encode, using the formalism defined in Section 2.1: knowledge from the GEM (LGEM$^+$Components **C1** to **C4**); the medium in which the yeast is growing, represented by axioms in the theory for the presence of compounds in the extracellular space (LGEM$^+$Component **C6**); the availability of ubiquitous compounds in each cellular compartment and the extracellular space (LGEM$^+$Component **C6**); and the presence and expression of genes (LGEM$^+$Component **C5**). Deduction can be used to analyse pathways and reachable metabolites. In the case of growth/no-growth simulations, G represents the availability of all the essential compounds in the cytoplasm (LGEM$^+$Component **C7**). So if $T \models G$ we say that there is growth, otherwise not. Other goals used here are the availability of other endpoints of biochemical pathways. T and G are provided to the ATP in plain text files and plain text proofs are output. The logical proofs (that the goal is reachable) found by the ATP correspond to predicted active biochemical pathways.

2.3 Pathway extraction

As mentioned in Section 2.2, the logical proofs that a given goal is reachable correspond to predicted active biochemical pathways in *S. cerevisiae* that produce the goal compounds. More specifically, during the subsumption stage of the proof output, clauses stating availability of reaction products are resolved by using the clauses representing reactions into the necessary reactants and gene products. Each stage of the proof resolves reaction products into reactants and gene products, back through the reaction network until axioms for compound availability and gene expression are reached. Thus the proof contains a possible (but not necessarily unique) set of reactions, gene products and starting compounds necessary to produce the goal.

This information defines a possible pathway configuration for the given experimental setup. Fig. 2 shows an example of the extracted proof a simple pathway, the glycolytic pathway, resulting in the production of pyruvate from glucose.

2.4 Single-gene Essentiality Prediction

Essential genes are those genes whose removal from the genome leads to a loss of viability for the organism. We predict single-gene essentiality for *S. cerevisiae*. Gene knockouts were performed by negating the gene presence axiom in the logical theory (i.e. $\mathsf{gn}(\mathsf{gene})$ becomes $\neg\mathsf{gn}(\mathsf{gene})$), and predicting growth/no-growth according to the method in Section 2.2.

We compare predictions against lists of viable and inviable strains from a genome-wide collection of single gene deletion mutants for *S. cerevisiae* using several media (Giaever et al, 2002). In particular, we compare with cultivations on a minimal

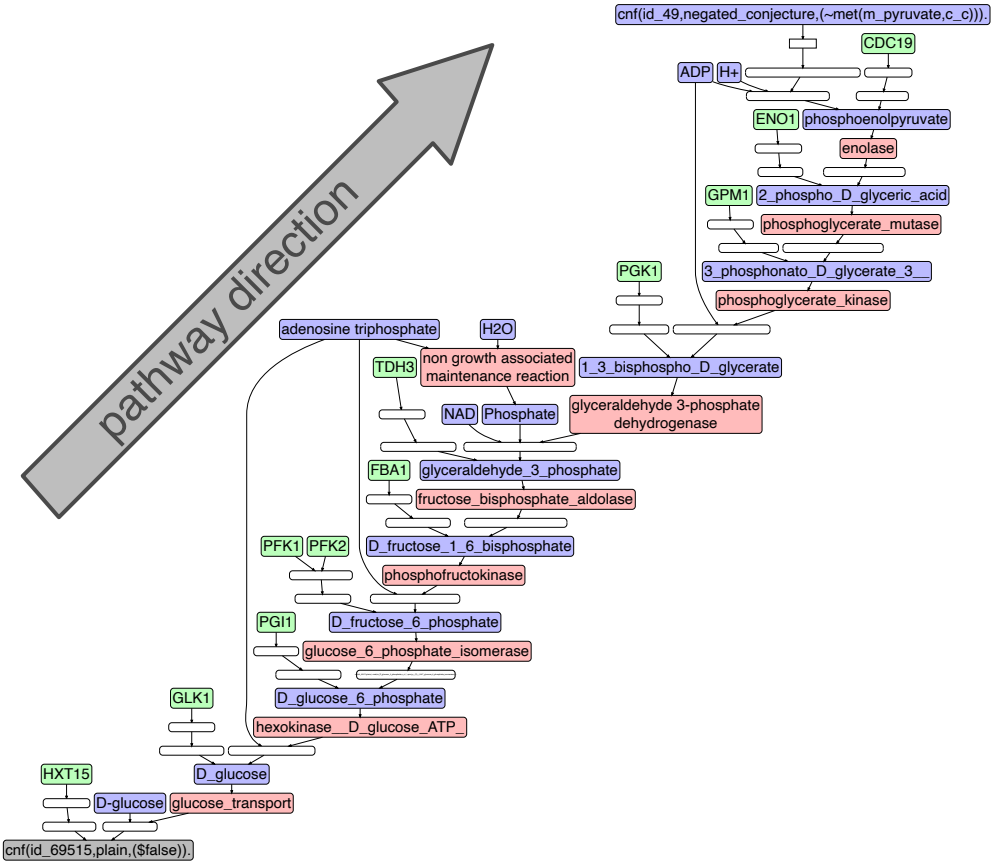

Fig. 2 Proof output from LGEM$^+$ having found a viable biochemical pathway for the production of pyruvate from glucose 6-phosphate (glycolysis). This pathway corresponds well with the literature. Given the query in the form of a negated conjecture, the proof takes the form of a refutation (top to bottom in this diagram). The resultant pathway is then from bottom to top. Green nodes represent expressed genes, blue nodes are intermediary metabolites, and red nodes represent activated reactions (the labels of these nodes have been simplified from the raw ATP output to improve readability). The remaining nodes represent intermediary steps in the refutation.

growth medium with the addition of uracil, histidine and leucine. A minimal growth medium is one containing a minimal subset of nutrients allowing for growth of the wild-type. The strain background used in this study was S288C, and in the experimental strain (*MATα/α his3D1/his3D1 leu2D0 /leu2D0 lys2D0/LYS2 MET15/met15D0 ura3D0 /ura3D0*) HIS3, LEU2, LYS2, MET17 and URA3 were not functional—for our experiments we remove these genes by default.

There are two basic error types with these predictions. We follow the naming convention as in Kumar and Maranas (2009), that we have: (1) *gNG inconsistency:* a prediction of growth when experimental data show no growth; and (2) *ngG inconsistency:* a prediction of no growth when experimental data show growth. Inconsistencies

arise from three main sources: deficiencies in the prior knowledge; errors in the prediction process; or conflicting empirical evidence. However it is the deficiencies in the prior knowledge that are of most interest for scientific discovery, which we explore next.

2.5 Using LGEM$^+$ Proofs to Constrain Flux Balance Analysis Simulations

Flux balance analysis (FBA) finds a reaction flux distribution, $\boldsymbol{\nu}$, given stoichiometric constraints from the GEM and a biologically relevant optimisation objective, $f(\boldsymbol{\nu})$, for example maximisation of biomass production (Orth et al, 2010; García Sánchez and Torres Sáez, 2014). FBA assumes the metabolism is in steady state, resulting in the constraint $S\boldsymbol{\nu} = \mathbf{0}$, where S is the stoichiometric matrix for the metabolic network and $\boldsymbol{\nu}$ is the reaction flux vector ($S \in \mathbb{Z}^{m \times n}$, where m is the number of compounds and n is the number of reactions in the metabolic network).

$$\begin{aligned} \underset{\boldsymbol{\nu} \in \mathbb{R}^n}{\text{maximize}} \quad & f(\nu_1, \ldots, \nu_n) \\ \text{subject to} \quad & S\boldsymbol{\nu} = \mathbf{0}, \\ & \nu_i^{\text{LB}} \leq \nu_i \leq \nu_i^{\text{UB}}, \quad i = 1, \ldots, n. \end{aligned}$$

Whilst the stoichiometric matrix is fixed, the upper and lower bounds for each reaction can be set to achieve relevant results. Existing methods to set these bounds include integrating experimental measurements of fluxes, or using enzyme turnover rates and availability (Domenzain et al, 2022). We use FBA to assess the feasibility of proofs found using the ATP by: setting reaction bounds [1] based on pathways activated in the proof, as described in Section 2.3; and then solving the resultant optimisation problem. We are able to do this neatly as both use the same GEM as the knowledge source. Pseudo-code for the above algorithm is provided in Algorithm 1.

There are two main applications of Algorithm 1 in LGEM$^+$. The first is to assess the feasibility of pathways of deletant strains during the single-gene essentiality prediction task; we discuss this in Section 3.4. The second is to filter hypotheses generated by the abduction algorithm, as stated in Algorithm 2, Line 18.

2.6 Abduction of Hypotheses

Abduction is used to suggest hypotheses that resolve inconsistencies between our model and empirical data. We select a reasonable set of candidate hypotheses through a two-stage process: firstly, we generate hypotheses; and secondly, we rank and filter these according to relevant scientific criteria which we describe in Section 2.6.2.

[1] Metabolite concentrations vary substantially between compounds, as do reaction fluxes, so finding a forcing threshold which is appropriate for all reactions is not straightforward. For our FBA simulations we used the Python package `cobrapy` (version 0.26.3) (Ebrahim et al, 2013); in the absence of relevant documentation on a suitable threshold, we found in a discussion for a MATLAB implementation of COBRA that a suitable threshold should be set at 1×10^{-9} mmol g_{DW}^{-1} h^{-1} (https://web.archive.org/web/20240327135202/https://groups.google.com/g/cobra-toolbox/c/9xmP1VcrWL0).

Algorithm 1 Constraining FBA solution given a logical theory T and a goal G

function FBACONSTRAIN(GEM, T, G, ν_0) ▷ ν_0 is minimum flux threshold for activation
  Use ATP to find proof of $T \models G$ ▷ The goal is reachable
  $i \leftarrow 1$
  for $i = 1..N$ **do** ▷ N is the number of reactions in the GEM
    if r_i active in the proof in the forward direction **then**
      $\nu_i^{LB} \leftarrow \nu_0$ ▷ Force reactions to have positive flux
    else if r_i active in the proof in the reverse direction **then**
      $\nu_i^{UB} \leftarrow -\nu_0$
    end if
  end for
  Solve FBA problem ($S\boldsymbol{\nu} = \mathbf{0}$) with resultant flux bounds
  return $(\boldsymbol{\nu}, \text{growthValue}, \text{solutionStatus}) \in \mathbb{R}^{\text{N}} \times \mathbb{R} \times \{\text{optimal}, \text{infeasible}\}$
end function

The abduction algorithm we designed for LGEM$^+$ is based on a loop through all genes in the model, and calculating hypotheses for each ngG error. Finally the hypotheses are filtered and ranked according to heuristics. Generating hypotheses using an ATP to fix a scientific theory is a general purpose technique that could be applied to other domains. Ranking and filtering heuristics will be domain-specific; here we describe the heuristics that we used, but others could well be applied. Pseudo-code for the abduction algorithm is provided in Algorithm 2. We now explain how different parts of this algorithm work.

2.6.1 Generating Candidate Hypotheses using an automated theorem prover

If the goal is not reachable (i.e. $T \not\models G$) the ATP abduces candidate hypotheses: sets H_i such that $\forall i$ $(T \wedge H_i \models G)$. This is done by reverse consequence finding $(T \wedge \neg G \models \neg H_i)$. For this project we extended iProver to include these features, which, not being specific to biochemical reaction networks, could be used for automated discovery in other scientific domains by constructing an appropriate FOL model. The form of the hypotheses, H_i, is a set of clauses expressed in terms of the predicates described above in Section 2.1. It is possible to restrict or guide the reverse consequence finding algorithm in iProver to seek certain types of hypotheses; we currently seek axioms that are all positive or all negative. For example a hypothesis could be: met(compound, compartment), that compound is available in compartment. Such hypotheses are challenging to discover because of the complexity of interaction in these networks.

Due to missing components in the models, none of the logical theories resultant from the conversion from Yeast8, iMM904 and iFF708 (the three GEMs used as background knowledge in this study, see Section 2.8) was viable given the minimal medium and ubiquitous compounds, even without any gene deletions, meaning one or more of the essential compounds was not produced. LGEM$^+$abduced hypotheses

consisting of combinations of compounds whose presence would enable viability of the base strain (e.g. for S288C, deletions for HIS3, LEU2, LYS2, MET17 and URA3). We chose the hypothesis with the fewest additional compounds.

For *ngG* inconsistencies there exists a set of essential metabolites not being produced that empirical data indicate will be produced given the specified genotype and conditions—in some sense the pathways in the model are incomplete. Hypotheses in this scenario are those that repair an incomplete pathway: additional reactions; annotation of an isoenzyme for knocked out genes; or removal of reaction annotations. For *gNG* inconsistencies there is a pathway in the model that empirical data suggest should be interrupted but is not. Thus hypotheses in this scenario will be those that interrupt a complete pathway: annotation of a pathway-critical reaction with a gene that is in the set of knocked out genes; removal of an isoenzyme annotation; or removal of reactions.

2.6.2 Heuristics for Ranking and Filtering Hypotheses

We filter hypotheses to only include either: (a) addition of one or more compounds (i.e. containing only atoms using the met predicate); or (b) the presence of one or more particular enzyme groups for a reaction (i.e. containing only atoms using the enz predicate). The motivation is that the subsequent model improvement step (to repair the pathway) for case (a) would be to add reactions to the model that produce the hypothesised metabolites, and for case (b) to either identify an isoenzyme for hypothesised groups or remove the annotation for the deleted gene for one of these reactions. We also remove hypotheses that introduced availability of one or more of the target compounds in the cytoplasm, as this would directly ensure the goal was reached but is of no scientific value.

We applied two criteria to assess the merit of each hypothesis. Firstly, by using our FBA constraint method, described in Section 2.5. Around half of the hypotheses resulted in infeasible solutions or very small growth—this means perhaps there might be something else that is missing from the model that is not included in the hypothesis. We filter out such hypotheses that incur infeasible FBA simulations. The second criteria was evaluating the impact each hypothesis had on the overall error in single-gene essentiality prediction. If the total number of *ngG* errors fixed is greater than the number of *gNG* errors introduced then this is a good hypothesis. Another, more conservative, approach would be to only add hypotheses to the model that do not introduce any *gNG* errors.

A final heuristic was whether hypotheses contained compounds that were not produced by any reaction in the GEM, meaning adding a suitable reaction that produces this compound would repair the error. These hypotheses could be tested experimentally by constructing a deletion mutant, cultivating with minimal medium and after observing growth, using metabolomic analysis (e.g. with mass spectrometry) to identify if the hypothesised intermediary metabolite set is present. If there were a reaction already in the GEM that produced the compound there could be other deficiencies in the model that need addressing first, for example gene annotation for those reactions. In this case LGEM$^+$ abduces hypotheses of case (b) above. Currently LGEM$^+$ can hypothesise to remove gene annotation, but this could be extended to include a

search for an isoenzyme based on similarity (e.g. sequence similarity) to the knocked out gene.

Algorithm 2 Abduction using LGEM$^+$

procedure AbductionSingleGene
  $\mathcal{H} \leftarrow \emptyset$
  for gene in all genes in theory **do**
    $T \leftarrow T$ ▷ Make a copy of the base theory
    $\widetilde{T} \leftarrow \widetilde{T} \setminus \{\mathsf{gn}(\mathsf{gene})\} \cup \{\neg\mathsf{gn}(\mathsf{gene})\}$ ▷ Construct deletant
    Use ATP to deduce if goal is reachable by identifying if $\widetilde{T} \vDash G$
    if $\widetilde{T} \vDash G$ **then** ▷ Growth prediction
      continue
    else if $\widetilde{T} \nvDash G$ **then** ▷ Non-growth prediction
      if gene is essential **then** ▷ No growth observed; no error
        continue
      else if gene is not essential **then** ▷ Growth observed; ngG error
        Abduction of potential hypotheses set $\mathcal{H}_{\mathsf{gene}}$ using ATP
        $\mathcal{H} \leftarrow \mathcal{H} \cup \mathcal{H}_{\mathsf{gene}}$
      end if
    end if
  end for
  Filter and rank $\mathcal{H} = \bigcup_{\mathsf{gene} \in \text{theory}} \mathcal{H}_{\mathsf{gene}}$, according to heuristics, e.g. Section 2.6
end procedure

2.7 Model-driven experiment design

Improvement of models is fuelled by comparing predictions against relevant experimental data. We use LGEM$^+$ to design experiments, and the resulting specification can be used to conduct the experiments, or query an experiment database. Online databases, such as ArrayExpress, contain experimental data from previous studiesAthar et al (2019).

The experimental factors over which LGEM$^+$ can vary are strain and growth medium. To identify the most interesting candidates for experimentation, we performed pathway simulations according to the methodology in Section 2.3 for candidate mutant strains (BY4742Δg, for a given gene g) in different media conditions. We selected deletant strains that displayed marked deviation in predicted expressed genes from the wild-type.

We queried ArrayExpress for a relevant study, and found one that obtained microarray expression data for wild-type (BY4742) and single-gene deletant strains (based on BY4742), grown in a glucose minimal medium (Kemmeren et al, 2014). LGEM$^+$ predicts a possible pathway configuration and which metabolic genes will be necessarily expressed, and we compare differential expression data from experiments with these predictions. The results of these comparisons are discussed in Section 3.5.

2.8 Sources of Knowledge

The primary source of the knowledge about reactions and associated genes is the GEM Yeast8 (v8.46.4.46.2) (Lu et al, 2019). This was chosen due to its broad coverage of the reactions and gene associations as well as its specificity to the organism *S. cerevisiae*. The other two GEMs used were: iMM904 (Mo et al, 2009) and iFF708 (Förster et al, 2003). (We include iFF708 as a background knowledge source partly to enable comparison with previous logical modelling approach (Whelan and King, 2008).) The models are stored using Systems Biology Markup Language (SBML). The software written to convert a GEM SBML file to a logical knowledge base is available in the supporting material, and follows the process described below and shown in Fig. 1.

We use three reference lists of compounds from Whelan and King (2008); these are shown in the first column of the files on the LGEM$^+$ GitHub repository[2] corresponding to: (1) all compounds deemed essential for growth in *S. cerevisiae*[3]; (2) compounds assumed ubiquitous during growth assumed to be present throughout the cell regardless of initial conditions, such as H_2O and O_2[4]; and (3) the growth media for the experiments, in this case yeast nitrogen base (YNB) with addition of ammonium, glucose and three amino acids (uracil, histidine and leucine)[5].

Each compound in these lists has an associated Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa, 2000) identifier. We matched compounds in the curated GEMs based firstly on KEGG ID, otherwise using the species name or synonyms. Some of the compounds we wish to include do not have corresponding entities in the GEMs used as background knowledge. Therefore there are discrepancies between the reference lists and the compiled lists.

3 Results

3.1 Automated Theorem Proving Efficiently Predicts Single-gene Essentiality

One of the purposes of LGEM$^+$ as outlined in Section 1 is to be able to efficiently predict phenotype for given genotype and conditions. To demonstrate this we address the task of single gene essentiality prediction.

Using three GEMs—Yeast8, iMM904 and iFF708—as background knowledge sources we conducted single-gene deletant simulations to assess essentiality of each gene and compared against a genome-wide deletion mutant cultivation (Giaever et al, 2002). Detailed descriptions of these methods are provided in Section 2. A summary of the single-gene essentiality prediction results is provided in Table 2.

When compared to previous qualitative methods our method showed superior results (Wunderlich and Mirny, 2006; Whelan and King, 2008). Yet quantitative prediction using FBA achieves a higher precision and recall. These error rates indicate how much is still to be learnt about yeast metabolism. We also found that gene essentiality predictions vary somewhat depending on the prior.

---

[2] https://github.com/AlecGower/LGEMPlus
[3] `src/model-files/essential-compounds-{model}.tsv`
[4] `src/model-files/ubiquitous-compounds-{model}.tsv`
[5] `src/model-files/ynb-compounds-{model}.tsv`

Base GEM	LGEM$^+$(Yeast8)	LGEM$^+$(iMM904)	LGEM$^+$(FF708)	FBA (Yeast8)	*Syn. Acc.* (Wunderlich and Mirny, 2006)
# predictions (#genes in GEM)	1056 (1150)	827 (905)	566 (619)	1068 (1150)	*682*
NG Recall (*ngNG/*NG*)	0.193 (31/161)	0.266 (33/124)	0.140 (14/100)	0.447 (72/161)	*0.119 (14/118)*
NG Precision (*ngNG/ng**)	0.431 (31/72)	0.478 (33/69)	0.778 (14/18)	0.459 (72/157)	*0.292 (14/48)*
gNG Rate (*gNG/*NG*)	0.807 (130/161)	0.734 (91/124)	0.860 (86/100)	0.553 (89/161)	*0.881 (104/118)*
ngG Rate (*ngG/*G*)	0.046 (41/895)	0.051 (36/703)	0.009 (4/466)	0.094 (85/907)	*0.060 (34/564)*
F1 score	0.266	0.342	0.237	0.453	*0.169*

Table 2 Comparative prediction results for single-gene essentiality using LGEM$^+$ across three background knowledge sources: Yeast8 (v8.46.4.46.2); iMM904; and iFF708, with comparison to: (a) an FBA-simulation with a viability threshold on growth rate set at 1×10^{-6} h^{-1} (according to Lu et al (2019)); and (b) another qualitative prediction method, the "synthetic accessibility" approach taken by Wunderlich et. al. (Wunderlich and Mirny, 2006). The empirical data used as truth data for these statistics were taken from a genome-wide screening study using a minimal medium (Giaever et al, 2002). The FOL model performance represents an improvement on previous qualitative method.
Shorthand: **NG*-observed no growth; **G*-observed growth; *ng**-predicted no growth. (Note that the performance statistics for the synthetic accessibility method are taken directly from the authors' report so there may be a difference in truth data to those used to evaluate our model.)

Simulation times for gene knockouts also appear to scale linearly with the size of the network. Comparing network size to average gene knockout simulation times for the three GEMs tested, we see that the mean (±1s.d.) times for one knockout simulation were: 0.52s±0.09s for iFF708 (1379 reactions); 0.67s±0.12s for iMM904 (1577 reactions); and 1.46s±0.32s for Yeast8 (4058 reactions).

3.2 Abductive Reasoning allows for Identification of Possible Missing Reactions

Another purpose of LGEM$^+$ we to abduce hypotheses to improve the model. Here we demonstrate the LGEM$^+$ abduction procedure on the Yeast8 model.

The first useful result obtained from abduction using LGEM$^+$ was a fix for the central carbon metabolism pathway. Deductive simulations showed that the pathway for central carbon metabolism was not complete, and that in particular production of alpha-D-Glucose 6-phosphate was not possible for the wildtype. We provided LGEM$^+$ with a list of possible compound synonyms. The predicted production of alpha-D-Glucose 6-phosphate for the wildtype, given the hypothesis that D-Glucose 6-phosphate was a synonym for alpha-D-Glucose 6-phosphate. This matches with experimental data, and so this hypothesis was adopted into the model.

Next we applied the abduction procedure to each of the 41 *ngG* errors in the single-gene deletion task. We generated candidate hypotheses according to methods described in Section 2.6. In total we generated 2094 unique hypotheses; some hypotheses would result in an error correction for several genes. We ranked and filtered these hypotheses according to domain-specific heuristics, finding 681 of these contained only met (633) or enz (48) predicates. We filtered to these 681 hypotheses because they correspond to changes that could be made to the model or in an experiment, e.g. by adding a compound to the growth medium or by modifying a gene protein reaction rule. The FBA evaluation outlined in Section 2.5 indicated 534 hypotheses that could be balanced by the reactions forced in the model. The overlap between these two sets was 118 hypotheses. For 14 of these hypotheses, adding them to the theory resulted

in a net improvement on the single-gene prediction task—i.e. they fixed more *ngG* errors than the number of *gNG* errors introduced.

3.3 Strict Essentiality Criteria and Incomplete Annotation may Explain *ngG* and *gNG* Inconsistencies

If just one essential compound is not produced we have no growth. One result of this setup is a relatively low precision in the single-gene essentiality prediction. Of the 72 deletions predicted inviable by our model, 41 of these are shown to result in experimentally viable mutant strains (*ngG* errors).

For several genes in the L-arginine biosynthesis pathway, the only essential metabolite in the model not reachable after deletion was L-arginine. These resulted in *ngG* errors despite the pathway structure and previous empirical evidence showing that null mutants for genes in this pathway (e.g. for *ARG1* (Crabeel et al, 1988)) are auxotrophic for L-arginine (i.e. L-arginine was not produced); in other words, there is conflicting experimental evidence. These cases are candidates for experimental testing, and highlight the potential of such models to inform laboratory experimental design and research direction.

In the Yeast8 model there are 4058 reactions, 1425 (35%) of which have no enzyme annotation and 540 (13%) are annotated with a set of isoenzymes that do not have a specific gene in common. Thus nearly half of all reactions will not be affected by single-gene deletions, which is likely to account for a portion of the 130 *gNG* inconsistencies in LGEM$^+$ single-gene essentiality predictions.

3.4 Pathways Output from LGEM$^+$ Overlap with FBA Simulations

In the case of a prediction of growth LGEM$^+$ outputs reaction pathways. FBA simulations output a reaction flux distribution, and from this we can use a flux threshold for reaction activate to obtain reaction pathways. When comparing reaction pathways obtained from both methods, for each deletant simulation just over 50% of reactions in the LGEM$^+$ derived pathways are also active in the FBA pathways. By comparison, around 30% of reactions in FBA derived pathways are also active in the LGEM$^+$ derived pathways. This indicates that there are pathways that are not on the critical path for production of essential compounds that are activated in FBA simulations.

Using pathways derived from the FBA constraint method described in Section 2.5, we investigated the *gNG* errors. Of the 130 errors, 50 of them resulted in pathways that the FBA method indicated were unfeasible (i.e., they resulted in low or zero growth). This would mean that by including this constraint method in the LGEM$^+$ framework we could eliminate these errors. However doing so would also falsely predict 56 viable deletant strains as inviable (new *ngG* errors).

3.5 Microarray expression results

Following the methodology for model-driven experimental design described in Section 2.7, we identified possible pathways, and associated gene expression profiles,

for single-gene deletant strains. Across all mutants in the experimental data set, the overlap between LGEM$^+$ and the microarray studyKemmeren et al (2014) was 113 strains, of which ten strains were predicted to have different expression profiles from the wildtype. For 62% of genes predicted to be differentially expressed in these strains by LGEM$^+$, the directionality of differential expression (up/down) was predicted correctly; applying McNemar's test against a predictor based on the 50% class balance, we acheived a p-value of 0.12. This statistic is not sufficient to reject the null hypothesis that the difference between our predictions and the majority class estimator are not significant.

We identified Δ*pfk2* (strain with the gene *PFK2* deleted from the genome) as one that showed a large deviation from the wild-type strain. *PFK2* is of particular interest as it is involved in the central carbon metabolism, an critical set of biochemical pathways for eukaryotes.

For Δ*pfk2* there were five genes predicted to be overexpressed (*TPI1*, *SOL3*, *GND1*, *TAL1*, and *ZWF1*) and one gene expected to be underexpressed (*FBA1*), when compared with the wild-type. *PFK2* encodes for the alpha subunit of phosphofructokinase-1. As shown in Fig. 3(a,b), we see that phosphofructokinase-1 is a required enzyme for the conversion of fructose 6-phosphate (F6P) into fructose 1,6-biphosphate (FBP) and onward to glyceraldehyde 3-phosphate (GAP) (Heillisch et al, 1989). In the absence of *PFK2*, an alternative pathway is available for the production of GAP, via the pentose phosphate pathway. The predicted expression changes correspond to such a switch.

We have reasoned about the feasibility of the prediction, but now we look to test this against experimental data. Thus we compare the predictions against microarray expression data for wild-type (BY4742) and Δ*pfk2* mutant (based on BY4742) (Kemmeren et al, 2014). Of those genes that LGEM$^+$ predicted would be overexpressed in the mutant strain, four were significantly so (*TPI1*, *GND1*, *TAL1*, and *ZWF1*); *SOL3* was not significantly differentially expressed. *FBA1*, which was predicted underexpressed, was significantly differentially expressed, however it was observed overexpressed rather than under.

As shown, the predictions for *PFK2* were reasonable given the background knowledge of the glycolytic pathway. And they corresponded with observed differential expression. The nature of the action of *PFK2* lead to a clear difference in a pathway. But overall there were many more genes differentially expressed that LGEM$^+$ predicted. One likely explanation for this is the role of gene regulation and signalling processes, which are not included in LGEM$^+$ here.

4 Discussion and Conclusion

Scientific discovery in biology is difficult due to the complexity of the systems involved and the expense of obtaining high quality experimental data. Automated techniques that make good use of background knowledge, of which GEMs are prime examples, will have a strong starting point. LGEM$^+$ seeks to do just that by using FOL combined with a powerful automated theorem prover (ATP).

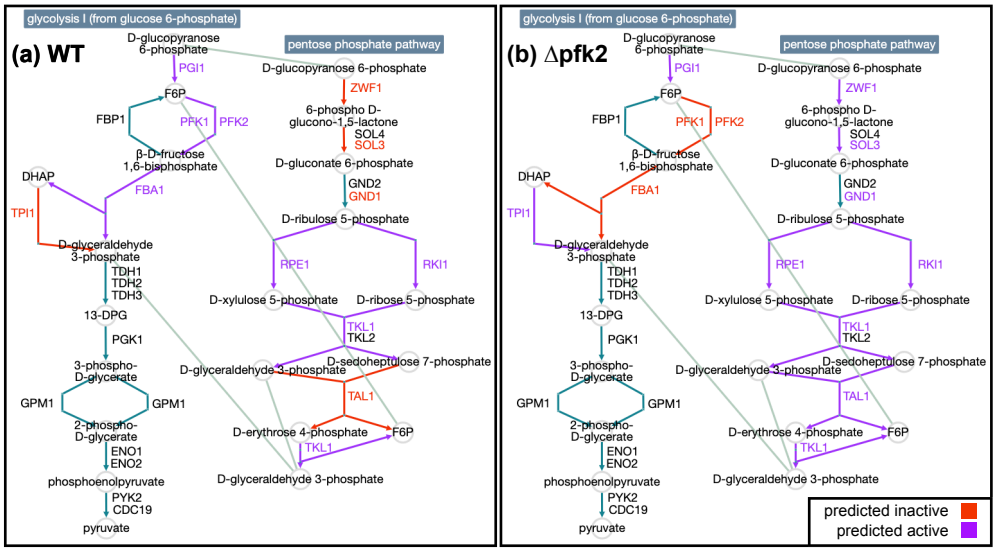

Fig. 3 **(a, b)** Predicted reconfiguration of the glycolytic pathway from wild-type **(a)** and after deletion of *PFK2* **(b)** to make use of the pentose phosphate pathway to produce glyceraldehyde 3-phosphate (GAP). Nodes represent molecules in the pathway, and edges are enzymatic reactions with associated genes. Purple reactions are those predicted active, and red those predicted not active.

We efficiently predicted single-gene essentiality in *S. cerevisiae* using a first-order logic (FOL) model. Our method showed superior results compared to previous qualitative methods, yet quantitative prediction using FBA achieves a higher precision and recall. We designed and implemented an algorithm for the abduction of hypotheses for improvement of a GEM. We found 633 hypotheses proposing availability of compounds in specific compartments, and therefore indicate possible missing reactions, 118 of which were validated through FBA constraint and 14 of which resulted in improvements in the single-gene essentiality prediction task. These heuristics help to select more promising hypotheses for experimentation; further selection will be informed by viability or cost of experiment design. We are integrating this with automated laboratories, including the robot scientist Genesis, which is based around chemostat cultivation and high-throughput metabolomics. As we scale the system we can adjust parameters in the heuristics, or introduce new heuristics, to return only the most promising hypotheses.

Measuring performance statistics relative to the number of genes in a model, rather than the number of genes in the organism, presents some challenges when designing a learning process to improve this performance (e.g. GrowMatch (Kumar and Maranas, 2009)). This highlights the need for better model assessment criteria to drive abduction. We have provided two examples here, namely: the constraint of FBA solutions; and comparison of pathway predictions with gene expression data from experiments. The best way to test hypotheses is through *in vivo* experimentation. We integrated LGEM$^+$ into a model-driven experimental design process. Future work could certainly

be directed to defining more criteria, particularly ones that exploit experimental data, and integrating them into LGEM$^+$.

The logical theory developed here was focused on efficient inference on biochemical pathways. A challenge for future development is to extend the first-order vocabulary to improve the power and performance of LGEM$^+$. Extending the vocabulary could mean: including more predicates, increasing the arity (number of arguments) of predicates, and introducing other logical clause forms. All to better encode biological processes, for example more detail regarding enzyme availability, integration of gene regulation and signalling or introducing time-dependent processes. Aligning the logic more closely with existing ontologies, for example the Systems Biology Ontology (SBO), would ensure the theory remains useful and semantically precise as it is extended. This is a common challenge across the scientific discovery community as we move further toward joint teams of human and robot scientists—ontologies provide a common language. Using FOL allows us to work toward connecting LGEM$^+$ with external knowledge bases.

Code and Data Availability: Code and data used in this study, including all model files for background knowledge, are available at https://github.com/AlecGower/LGEMPlus. Logical inference was performed using iProver (v3.7).

Statements and Declarations: The authors have no competing interests to declare that are relevant to the content of this article.

References

Athar A, Füllgrabe A, George N, et al (2019) ArrayExpress update – from bulk to single-cell expression data. Nucleic Acids Research 47(D1):D711–D715. https://doi.org/10.1093/nar/gky964, URL https://academic.oup.com/nar/article/47/D1/D711/5144130

Chen Y, Li F, Nielsen J (2022) Genome-scale modeling of yeast metabolism: Retrospectives and perspectives. FEMS Yeast Research 22(1):foac003. https://doi.org/10.1093/femsyr/foac003

Crabeel M, Seneca S, Devos K, et al (1988) Arginine repression of the Saccharomyces cerevisiae ARG1 gene. Comparison of the ARG1 and ARG3 control regions. Current Genetics 13(2):113–124. https://doi.org/10.1007/BF00365645

Domenzain I, Sánchez B, Anton M, et al (2022) Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0. Nature Communications 13(1):3766. https://doi.org/10.1038/s41467-022-31421-1

Dujon B (2010) Yeast evolutionary genomics. Nature Reviews Genetics 11(7):512–524. https://doi.org/10.1038/nrg2811

Ebrahim A, Lerman JA, Palsson BO, et al (2013) COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC Systems Biology 7(1):74. https://doi.org/10.1186/1752-0509-7-74

Förster J, Famili I, Fu P, et al (2003) Genome-Scale Reconstruction of the Saccharomyces cerevisiae Metabolic Network. Genome Research 13(2):244–253. https://doi.org/10.1101/gr.234503

García Sánchez CE, Torres Sáez RG (2014) Comparison and analysis of objective functions in flux balance analysis. Biotechnology Progress 30(5):985–991. https://doi.org/10.1002/btpr.1949

Giaever G, Chu AM, Ni L, et al (2002) Functional profiling of the Saccharomyces cerevisiae genome. Nature 418(6896):387–391. https://doi.org/10.1038/nature00935

Goffeau A, Barrell BG, Bussey H, et al (1996) Life with 6000 Genes. Science 274(5287):546–567. https://doi.org/10.1126/science.274.5287.546

Gower AH, Korovin K, Brunnsåker D, et al (2023) LGEM$^+$: A First-Order Logic Framework for Automated Improvement of Metabolic Network Models Through Abduction. In: Bifet A, Lorena AC, Ribeiro RP, et al (eds) Discovery Science. Springer Nature Switzerland, Cham, pp 628–643

Heavner BD, Price ND (2015) Comparative Analysis of Yeast Metabolic Network Models Highlights Progress, Opportunities for Metabolic Reconstruction. PLOS Computational Biology 11(11):e1004530. https://doi.org/10.1371/journal.pcbi.1004530

Heillisch J, Ritzel R, von Borstel R, et al (1989) The phosphofructokinase genes of yeast evolved from two duplication events. Gene 78(2):309–321. https://doi.org/10.1016/0378-1119(89)90233-3, URL http://dx.doi.org/10.1016/0378-1119(89)90233-3

Kanehisa M (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 28(1):27–30. https://doi.org/10.1093/nar/28.1.27

Kemmeren P, Sameith K, van de Pasch LA, et al (2014) Large-Scale Genetic Perturbations Reveal Regulatory Networks and an Abundance of Gene-Specific Repressors. Cell 157(3):740–752. https://doi.org/10.1016/j.cell.2014.02.054

King RD, Whelan KE, Jones FM, et al (2004) Functional genomic hypothesis generation and experimentation by a robot scientist. Nature 427(6971):247–252. https://doi.org/10.1038/nature02236

Korovin K (2008) iProver – An Instantiation-Based Theorem Prover for First-Order Logic (System Description). In: Armando A, Baumgartner P, Dowek G (eds) Automated Reasoning, vol 5195. Springer Berlin Heidelberg, Berlin, Heidelberg, p 292–298, https://doi.org/10.1007/978-3-540-71070-7_24

Kumar VS, Maranas CD (2009) GrowMatch: An Automated Method for Reconciling In Silico/In Vivo Growth Predictions. PLoS Computational Biology 5(3):e1000308.

https://doi.org/10.1371/journal.pcbi.1000308

Lu H, Li F, Sánchez BJ, et al (2019) A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. Nature Communications 10(1). https://doi.org/10.1038/s41467-019-11581-3

Mo ML, Palsson B, Herrgård MJ (2009) Connecting extracellular metabolomic measurements to intracellular flux states in yeast. BMC Systems Biology 3. https://doi.org/10.1186/1752-0509-3-37

Orth JD, Thiele I, Palsson BØ (2010) What is flux balance analysis? Nature Biotechnology 28(3):245–248. https://doi.org/10.1038/nbt.1614

Prigent S, Frioux C, Dittami SM, et al (2017) Meneco, a Topology-Based Gap-Filling Tool Applicable to Degraded Genome-Wide Metabolic Networks. PLOS Computational Biology 13(1):e1005276. https://doi.org/10.1371/journal.pcbi.1005276

Reiser PGK, King RD, Muggleton SH, et al (2001) Developing a logical model of yeast metabolism. Electronic Transactions in Artificial Intelligence 5(B):223–244

Rozanski R, Bragaglia S, Ray O, et al (2015) Automating the Development of Metabolic Network Models. In: Roux O, Bourdon J (eds) Computational Methods in Systems Biology. Springer International Publishing, Cham, Lecture Notes in Computer Science, pp 145–156, https://doi.org/10.1007/978-3-319-23401-4_13

Simon L, del Val A (2001) Efficient consequence finding. In: Nebel B (ed) Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, IJCAI 2001, Seattle, Washington, USA, August 4-10, 2001. Morgan Kaufmann, pp 359–370

Tamaddoni-Nezhad A, Chaleil R, Kakas A, et al (2005) Abduction and induction for learning models of inhibition in metabolic networks. In: Arif Wani M, Milanova M, Kurgan L, et al (eds) Fourth International Conference on Machine Learning and Applications (ICMLA'05), pp 233–238, https://doi.org/10.1109/ICMLA.2005.6

Tamura T, Lu W, Akutsu T (2015) Computational Methods for Modification of Metabolic Networks. Computational and Structural Biotechnology Journal 13:376–381. https://doi.org/10.1016/j.csbj.2015.05.004

Thiele I, Palsson BØ (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature Protocols 5(1):93–121. https://doi.org/10.1038/nprot.2009.203

Whelan KE, King RD (2008) Using a logical model to predict the growth of yeast. BMC bioinformatics 9:97. https://doi.org/10.1186/1471-2105-9-97

Wunderlich Z, Mirny LA (2006) Using the Topology of Metabolic Networks to Predict Viability of Mutant Strains. Biophysical Journal 91(6):2304–2311. https://doi.org/

10.1529/biophysj.105.080572